Data Analysis with Excel®

An Introduction for Physical Scientists

Les Kirkup
University of Technology, Sydney

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

© L. Kirkup 2002

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2002
Reprinted 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Utopia 9.25/13.5pt. *System* QuarkXPress® [SE]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Kirkup, Les.
Data analysis with Excel : an introduction for physical scientists / Les Kirkup.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-79337-8 – ISBN 0-521-79737-3 (pb.)
1. Research–Statistical methods–Data processing. 2. Electronic spreadsheets.
3. Microsoft Excel for Windows. I. Title.

Q180.55.S7 K57 2002
001.4′22′0285–dc21 2001037408

ISBN 0 521 79337 8 hardback
ISBN 0 521 79737 3 paperback

Contents

Chapter 1

Introduction to scientific data analysis

1.1 Introduction

> 'The principle of science, the definition almost, is the following: *The test of all knowledge is experiment.* Experiment is the *sole judge* of scientific "truth"'.
>
> So wrote Richard Feynman, famous scientist and Nobel prize winner, noted for his contributions to physics.[1]

It is possible that when Feynman wrote these words he had in mind elaborate experiments devised to reveal the 'secrets of the universe', such as those involving the creation of new particles during high energy collisions in particle accelerators. However, experimentation encompasses an enormous range of more humble (but extremely important) activities such as testing the temperature of a baby's bath water by immersing an elbow into the water, or pressing on a bicycle tyre to establish whether it has gone 'flat'. The absence of numerical measures of quantities most distinguishes these experiments from those normally performed by scientists.

Many factors directly or indirectly influence the fidelity of data gathered during an experiment such as the quality of the experimental design, experimenter competence, instrument limitations and time available to perform the experiment. Appreciating and, where possible, accounting for such factors are key tasks that must be carried out by an experimenter. After every care has been taken to acquire the best data possible, it is time to apply techniques of data analysis to extract the most from the data. The

[1] See Feynman, Leighton and Sands (1963).

process of extraction requires qualitative as well as quantitative methods of analysis. The first steps require consideration be given to how data may be summarised numerically and graphically and this is the main focus of this chapter.[2] Some of the ideas touched upon in this chapter, such as those relating to error and uncertainty, will be revisited in more detail in later chapters.

1.2 Scientific experimentation

To find out something about the world, we experiment. A child does this naturally, with no training or scientific apparatus. Through a potent combination of curiosity and trial and error, a child quickly creates a viable model of the 'way things work'. This allows the consequences of a particular action to be anticipated. Curiosity plays an equally important role in the professional life of a scientist who may wish to know:

- the amount of contaminant in a pharmaceutical;
- the thickness of the ozone layer in the atmosphere;
- the surface temperature of a distant star;
- the stresses experienced by the wings of an aircraft;
- the blood pressure of a person;
- the frequency of electrical signals generated within the human brain.

In particular, scientists look for relationships between quantities. For example, a scientist may wish to establish how the amount of energy radiated from a body each second depends on the temperature of that body. In formulating the problem, designing and executing the experiment and analysing the results, the intention may be to extend the domain of applicability of an established theory, or to present strong evidence of the breakdown of that theory. Where results obtained conflict with accepted ideas or theories, a key goal is to provide an alternative and better explanation of the results. Before 'going public' with a new and perhaps controversial explanation, the scientist needs to be confident in the data gathered and the methods used to analyse those data. This requires that experiments be well designed. In addition, good experimental design helps anticipate difficulties that may occur during the execution of the experiment and encourages the efficient use of resources.

Successful experimentation is often a combination of good ideas, good planning, perseverance and hard work. Though it is possible to dis-

[2] This is sometimes referred to as 'exploratory data analysis'.

cover something interesting and new 'by accident', it is usual for science to progress by small steps. The insights gained by researchers (both experimentalists and theorists) combine to provide answers and explanations to some questions, and in the process create new questions that need to be addressed. In fact, even if something new *is* found by chance, it is likely that the discovery will remain a curiosity until a serious scientific investigation is carried out to determine if the discovery or effect is real or illusory. While scientists are excited by new ideas, a healthy amount of scepticism remains until the ideas have been subjected to serious and sustained scrutiny by others.

Though it is possible to enter a laboratory with only a vague notion of how to carry out a scientific investigation, there is much merit in planning ahead as this promotes the efficient use of resources, as well as revealing whether the investigation is feasible or overambitious.

1.2.1 Aim of an experiment

An experiment needs a focus, more usually termed an 'aim', which is something the experimenter returns to during the design and analysis phases of the experiment. Essentially the aim embodies a question which can be expressed as 'what are we trying to find out by performing the experiment?'

Expressing the aim clearly and concisely at the outset is important, as it is reasonable to query as the experiment progresses whether the steps taken are succeeding in addressing the aim, or whether the experiment has deviated 'off track'. Heading off on a tangent from the main aim is not necessarily a bad thing. After all, if you observe an interesting and unexpected effect during the course of an experiment, it would be quite natural to want to know more, as rigidly pursuing the original aim might cause you to bypass an important discovery. Nevertheless, it is likely that if a new effect has been observed, this effect deserves its own separate and carefully planned experiment.

Implicit in the aim of the experiment is an idea or hypothesis that the experimenter wishes to promote or test, or an important question that requires clarification. Examples of questions that might form the basis of an experiment include:

- Is a new spectroscopic technique better able to detect impurities in silicon than existing techniques?
- Does heating a glass substrate during vacuum deposition of a metal improve the quality of the thin films deposited onto the substrate?

- To what extent does a reflective coating on windows reduce the heat transfer into a motor vehicle?
- In what way does the cooling efficiency of a thermoelectric cooler depend on the amount of electrical current supplied to the cooler?

Such questions can be restated explicitly as aims of a scientific investigation. It is possible to express those aims in a number of different, but essentially equivalent, ways. For example:

(a) The aim of the experiment is to determine the change in heat transfer to a motor vehicle when a reflective coating is applied to the windows of that vehicle.
(b) The aim of the experiment is to test the hypothesis that a reflective coating applied to the windows of a motor vehicle reduces the amount of heat transferred into that vehicle.

Most physical scientists and engineers would recognise (a) as a familiar way in which an aim is expressed in their disciplines. By contrast, the explicit inclusion of a hypothesis to be tested, as stated in (b), is often found in studies in the biological, medical and behavioural sciences. The difference in the way the aim is expressed is largely due to the conventions adopted by each discipline, as all have a common goal of advancing understanding and knowledge through experimentation and observation.

1.2.2 Experimental design

Deciding the aim or purpose of an experiment 'up front' is important, as precious resources (including the time of the experimenter) are to be devoted to the experiment. Experimenting is such an absorbing activity that it is possible for the aims of an experiment to become too ambitious. For example, the aim of an experiment might be to determine the effect on the thermal properties of a ceramic when several types of atoms are substituted for (say) atoms of calcium in the ceramic. If a month is available for the study, careful consideration must be given to the number of samples of ceramic that can be prepared and tested and whether a more restricted aim, perhaps concentrating on the substitution of just one type of atom, might not be more appropriate.

Once the aim of an experiment is decided, a plan of how that aim might be achieved is begun. Matters that must be considered include:

- What quantities are to be measured during the experiment?
- Over what ranges should the controllable quantities be measured?

- What are likely to be the dominant sources of error?
- What equipment is needed and what is its availability?
- In what ways are the data to be analysed?
- Does the experimenter need to become skilled at new techniques (say, how to operate an electron microscope, or perform advanced data analysis) in order to complete the experiment?
- Does new apparatus need to be designed/constructed/acquired or does existing equipment require modification?
- Is there merit in developing a computer based acquisition system to gather the data?
- How much time is available to carry out the experiment?
- Are the instruments to be used performing within their specifications?

A particularly important aspect of experimentation is the identification of influences that can affect any result obtained through experiment or observation. Such influences are regarded as sources of 'experimental error' and we will have cause to consider these in this text. In the physical sciences, many of the experimental variables that would affect a result are easily identifiable and some are under the control of the experimenter. Identifying sources that would adversely influence the outcomes of an experiment may lead to ways in which the influence might be minimised. For example, the quality of a metal film deposited onto a glass slide may be dependent upon the temperature of the slide during the deposition process. By improving the temperature control of the system, so that the variability of the temperature of the slide is reduced to (say) less than 5 °C, the quality of the films may be enhanced.

Despite the existence of techniques that allow us to draw out much from experimental data, a good experimenter does not rely on data analysis to 'make up' for data of dubious worth. If large scatter is observed in data, a sensible option is to investigate whether improved experimental technique can reduce the scatter. For example, time spent constructing electromagnetic shielding for a sensitive electronic circuit in an experiment requiring the measurement of extremely small voltages can improve the quality of the data dramatically and is to be much preferred to the application of 'advanced' data analysis techniques which attempt to compensate for shortcomings in the data.

An essential feature of experiments in the physical sciences is that the measurement process yields numerical values for quantities such as temperature, pH, strain, pressure and voltage. These numerical values (often referred to as *experimental data*) may be algebraically manipulated, graphed, compared with theoretical predictions or related to values

obtained by other experimenters who have performed similar experiments.

1.3 Units and standards

Whenever a value is recorded in a table or plotted on a graph, the unit of measurement must be stated, as numbers by themselves have little meaning. To encompass all quantities that we might measure during an experiment, we need units that are:

- comprehensive,
- clearly defined,
- internationally accepted,
- easy to use.

Reliable and accurate standards based on the definition of a unit must be available so that instruments designed to measure specific quantities may be compared with those standards. Without agreement between experimenters in, say, Australia and the United Kingdom as to what constitutes a metre or a second, a comparison of values obtained by each experimenter would be impossible.

A variety of instruments may be employed to measure quantities in the physical sciences, ranging from a 'low tech.' manometer to determine pressure in a chamber to a state of the art HPLC[3] to accurately determine the concentration of contaminant in a pharmaceutical. Whatever the particular details of a scientific investigation, we generally attach much importance to the 'numbers' that emerge from an experiment as they may provide support for a new theory of the origin of the universe, assist in monitoring damage to the earth's atmosphere or help save a life. Referring to the outcome of a measurement as a 'number' is rather vague and misleading. Through experiment we obtain *values*. A value is the product of a number and the unit in which the measurement is made. The distinction in scientific contexts between number and value is important. Table 1.1 includes definitions of number, value and other important terms as they are used in this text.

[3] HPLC stands for high performance liquid chromatography

Table 1.1. *Definitions of commonly used terms in data analysis.*

Term	Definition
Quantity	An attribute or property of a body, phenomenon or material. Examples of quantities are: the temperature, mass or electrical capacitance of a body; the time elapsed between two events such as starting and stopping a stop watch; and the resistivity of a metal.
Unit	An amount of a quantity, suitably defined and agreed internationally, against which some other amount of the same quantity may be compared. As examples, the kelvin is a unit of temperature, the second is a unit of time and the ohm-metre is a unit of resistivity.
Value	The product of a number and a unit. As examples, 273 K is a value of temperature, 0.015 s is a value of time interval and $1.7 \times 10^{-8}\ \Omega\cdot$m is a value of resistivity.
Measurement	A process by which a value of a quantity is determined. For example, the measurement of water temperature using an alcohol-in-glass thermometer entails immersing a thermometer in the water followed by estimating the position of the top of a narrow column of alcohol against an adjacent scale.
Data	Values obtained through measurement or observation.

1.3.1 Units

The most widely used system of units in science is the SI system[4] which has been adopted officially by most countries around the world. Despite strongly favouring SI units in this text, we will also use some 'non-SI units' such as the minute and the degree, as these are likely to remain in widespread use in science for the foreseeable future.

The origins of the SI system can be traced to pioneering work done on units in France in the late eighteenth century. In 1960 the name 'SI system' was adopted and at that time it consisted of six fundamental or 'base' units. Since 1960 the system has been added to and refined and remains constantly under review. From time to time suggestions are made regarding how the definition of a unit may be improved. If this allows for easier or more accurate realisation of the unit as a standard (permitting, for

[4] SI stands for Système International.

Table 1.2. *SI base units, symbols and definitions.*

Quantity	Unit	Symbol	Definition
Mass	kilogram	kg	The kilogram is equal to the mass of the international prototype of the kilogram. (The prototype kilogram is made from an alloy of platinum and iridium and is kept under very carefully controlled environmental conditions near Paris.)
Length	metre	m	The metre is the length of the path travelled by light in a vacuum during a time interval of $\frac{1}{299\,792\,458}$ of a second.
Time	second	s	The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.
Thermodynamic temperature	kelvin	K	The kelvin is the fraction $\frac{1}{273.16}$ of the thermodynamic temperature of the triple point of water.
Electric current	ampere	A	The ampere is that current which, if maintained between two straight parallel conductors of infinite length, of negligible cross-section and placed 1 metre apart in a vacuum, would produce between these conductors a force of 2×10^{-7} newton per metre of length.
Luminous intensity	candela	cd	The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{14} hertz and that has a radiant intensity in that direction of $\frac{1}{683}$ watt per steradian.
Amount of substance	mole	mol	The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12.

example, improvements in instrument calibration), then appropriate modifications are made to the definition of the unit. Currently the SI system consists of 7 base units as defined in table 1.2.

Other quantities may be expressed in terms of the base units. For example, energy can be expressed in units $\mathrm{kg{\cdot}m^2{\cdot}s^{-2}}$ and electric potential difference in units $\mathrm{kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-1}}$. The cumbersome nature of units expressed in this manner is such that other, so called *derived*, units are introduced which are formed from products of the base units. Some famil-

Table 1.3. *Symbols and units of some common quantities.*

Quantity	Derived unit	Symbol	Unit of quantity expressed in base units
Energy, work	joule	J	$kg{\cdot}m^2{\cdot}s^{-2}$
Force	newton	N	$kg{\cdot}m{\cdot}s^{-2}$
Power	watt	W	$kg{\cdot}m^2{\cdot}s^{-3}$
Potential difference, electromotive force (emf)	volt	V	$kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-1}$
Electrical charge	coulomb	C	$s{\cdot}A$
Electrical resistance	ohm	Ω	$kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-2}$

iar quantities with their units expressed in derived and base units are shown in table 1.3.

Example 1

The farad is the SI derived unit of electrical capacitance. With the aid of table 1.3, express the unit of capacitance in terms of the base units, given that the capacitance, C, may be written

$$C = \frac{Q}{V} \tag{1.1}$$

where Q represents electrical charge and V represents potential difference.

ANSWER

From table 1.3, the unit of charge expressed in base units is $s{\cdot}A$ and the unit of potential difference is $kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-1}$. It follows that the unit of capacitance can be expressed with the aid of equation (1.1) as

$$\frac{s \cdot A}{kg \cdot m^2 \cdot s^{-3} \cdot A^{-1}} = kg^{-1} \cdot m^{-2} \cdot s^4 \cdot A^2$$

Exercise A

The henry is the derived unit of electrical inductance in the SI system of units. With the aid of table 1.3, express the unit of inductance in terms of the base units, given the relationship

$$E = -L\frac{dI}{dt} \tag{1.2}$$

where E represents emf, L represents inductance, I represents electric current and t represents time.

1.3.2 Standards

How do the definitions of the SI units in table 1.2 relate to measurements made in a laboratory? For an instrument to measure a quantity in SI units, the definitions need to be made 'tangible' so that an example or *standard* of the unit is made available. Only when the definition is realised as a practical and maintainable standard can values obtained by an instrument designed to measure the quantity be compared against that standard. If there is a difference between the standard and the value indicated by the instrument, then the instrument is adjusted or *calibrated* so that the difference is minimised.

Accurate standards based on the definitions of some of the units appearing in table 1.2 are realised in specialist laboratories. For example, a clock based on the properties of caesium atoms can reproduce the second to high accuracy.[5] By comparison, creating an accurate standard of the ampere based directly on the definition of the ampere appearing in table 1.2 is much more difficult. In this case it is common for laboratories to maintain standards of related derived SI units such as the volt and the ohm, which can be implemented to very high accuracy.

Most countries have a 'national standards laboratory' which maintains the most accurate standards achievable, referred to as *primary* standards. From time to time the national laboratory compares those standards with other primary standards held in laboratories around the world. In addition, a national laboratory creates and calibrates secondary standards by reference to the primary standard. Such secondary standards are found in some government, industrial and university laboratories. Secondary standards in turn are used to calibrate and maintain working standards and eventually a working standard may be used to calibrate (for example) a hand held voltmeter used in an experiment. If the calibration process is properly documented, it is possible to trace the calibration of an instrument back to the primary standard.[6] 'Traceability' is very important in some situations, particularly when the 'correctness' of a value indicated by an instrument is in dispute.

1.3.3 Prefixes and scientific notation

Values obtained through experiment are often much larger or much smaller than the base (or derived) SI unit in which the value is expressed.

[5] See appendix 2 of The International System of Units (English translation) 7th Edition, 1997, published by the Bureau International des Poids et Mesures (BIPM).

[6] See Morris (1997), chapter 3.

Table 1.4. *Prefixes used with the SI system of units.*

Factor	Prefix	Symbol	factor	Prefix	Symbol
10^{-24}	yocto	y	10^{1}	deka	da
10^{-21}	zepto	z	10^{2}	hecto	h
10^{-18}	atto	a	**10^{3}**	**kilo**	**k**
10^{-15}	femto	f	**10^{6}**	**mega**	**M**
10^{-12}	**pico**	**p**	**10^{9}**	**giga**	**G**
10^{-9}	**nano**	**n**	**10^{12}**	**tera**	**T**
10^{-6}	**micro**	**μ**	10^{15}	peta	P
10^{-3}	**milli**	**m**	10^{18}	exa	E
10^{-2}	centi	c	10^{21}	zetta	Z
10^{-1}	deci	d	10^{24}	yotta	Y

In such situations there are two widely used methods by which the value of the quantity may be specified. The first is to choose a multiple of the unit and indicate that multiple by assigning a *prefix* to the unit. So, for example, we might express the value of the capacitance of a capacitor as 47 μF. The symbol μ stands for the prefix 'micro' which represents a factor of 10^{-6}. A benefit of expressing a value in this way is the conciseness of the representation. A disadvantage is that many prefixes are required in order to span the orders of magnitude of values that may be encountered in experiments. As a result, several unfamiliar prefixes exist. For example, the size of the electrical charge carried by an electron is about 160 zC. Only dedicated students of the SI system would immediately recognise z as the symbol for the prefix 'zepto' which represents the factor 10^{-21}. Table 1.4 includes the prefixes currently used in the SI system. The prefixes shown in bold are the most commonly used.

Another way of expressing the value of a quantity is to give the number that precedes the unit in scientific notation. To express any number in scientific notation, we separate the first non-zero digit from the second digit by a decimal point, so for example, the number 1200 becomes 1.200. So that the number remains unchanged we must multiply 1.200 by 10^3 so that 1200 is written as 1.200×10^3. Scientific notation is preferred for very large or very small numbers. For example, the size of the charge carried by the electron is written as 1.60×10^{-19} C. Though any value may be expressed using scientific notation, we should avoid taking this approach to extremes. For example, suppose the mass of a body is 1.2 kg. This *could* be written as 1.2×10^0 kg, but this is possibly going too far.

Example 2

Rewrite the following values using: (a) commonly used prefixes and (b) scientific notation:

(i) 0.012 s; (ii) 601 A; (iii) 0.00064 J.

ANSWER

(i) 12 ms or 1.2×10^{-2} s; (ii) 0.601 kA or 6.01×10^{2} A; (iii) 0.64 mJ or 6.4×10^{-4} J.

Exercise B

1. Rewrite the following values using prefixes:

   (i) 1.38×10^{-20} J in zeptojoules; (ii) 3.6×10^{-7}s in microseconds; (iii) 43258 W in kilowatts; (iv) 7.8×10^{8} m/s in megametres per second.

2. Rewrite the following values using scientific notation:

   (i) 0.650 nm in metres; (ii) 37 pC in coulombs; (iii) 1915 kW in watts; (iv) 125 μs in seconds.

1.3.4 Significant figures

In a few situations, a value obtained in an experiment can be exact. For example, in an experiment to determine the wavelength of light using Newton's rings,[7] the number of rings can be counted exactly. By contrast, the temperature of an object cannot be known exactly and so we must be careful when we interpret values of temperature. Presented with the statement that '*the temperature of the water bath was 21°C*' it is unreasonable to infer that the temperature was 21.0000000 °C. It is more likely that the temperature of the water was closer to 21 °C than it was to either 20 °C or 22 °C. By writing the temperature as 21 °C, the implication is that the value of temperature obtained by a single measurement is known to two figures, often referred to as *two significant figures*.

Inferring how many figures are significant simply by the way a number is written can sometimes be difficult. If we are told that the mass of a body is 1200 kg, how many figures are significant? If the instrument measures mass to the nearest 100 kg, then the 'real' mass lies between 1150 kg and 1250 kg, so in fact only the first two figures are significant. On

[7] See Smith and Thomson (1988).

the other hand, if the measuring instrument is capable of measuring to the nearest kilogram, then all four figures are significant. The ambiguity can be eliminated if we express the value using scientific notation. If the mass of the body, m, is correct to two significant figures we would write

$$m = 1.2 \times 10^3 \text{ kg}$$

When a value is written using scientific notation, every figure preceding the multiplication sign is regarded as significant. If the mass is correct to four significant figures then we write

$$m = 1.200 \times 10^3 \text{ kg}$$

Though it is possible to infer something about a value by the way it is written, it is better to state explicitly the uncertainty in a value. For example, we might write

$$m = (1200 \pm 12) \text{ kg}$$

where 12 kg is the uncertainty in the value of the mass. Estimating uncertainty is considered in chapter 5.

It may be required to round a value to a specified number of significant figures. For example, we might want to round 1.752×10^{-7} m to three significant figures. To do this, we consider the fourth significant figure (which in this example is a '2'). If this figure is equal to or greater than 5, we increase the third significant figure by 1, otherwise we leave the figure unchanged. So, for example, 1.752×10^{-7} m becomes 1.75×10^{-7} m to three significant figures. Using the same convention, a mass of 3.257×10^3 kg becomes 3.3×10^3 kg to two significant figures.

Exercise C

1. How many significant figures are implied by the way each of the following values is written:

(i) 1.72 m; (ii) 0.00130 mol/cm^3; (iii) 6500 kg; (iv) 1.701×10^{-3} V; (v) 100 °C; (vi) 100.0 °C?

2. Express the following values using scientific notation to two, three and four significant figures:

(i) 775710 m/s^2; (ii) 0.001266 s; (iii) –105.4 °C; (iv) 14000 nH in henrys; (v) 12.400 kJ in joules; (vi) 101.56 nm in metres

1.4 Picturing experimental data

The ability possessed by humans to recognise patterns and trends is so good that it makes sense to exploit this talent when analysing experimental data. Though a table of experimental values may contain the same information as appears on a graph, it is very difficult to extract useful information from a table 'by eye'. To appreciate the 'big picture' it is helpful to devise ways of graphically representing the values.

When values are obtained through repeat measurements of a single quantity, then the histogram is used extensively to display data. When a single quantity or variable is being considered, the data obtained are often referred to as 'univariate' data. By contrast, if an experiment involves investigating the relationship between two quantities, then the x–y graph is a preferred way of displaying the data (such data are often referred to as 'bivariate' data).

1.4.1 Histograms

The histogram is a pictorial representation of data which is regularly used to reveal the scatter or distribution of values obtained from measurements of a single quantity. For example, we might measure the diameter of a wire many times in order to know something of the variation of the diameter along the length of the wire. A table is a convenient and compact way to present the numerical information. However, we are usually happy (at least in the early stages of analysis) to forego knowledge of individual values in the table for a broader overview of all the data. This should help indicate whether some values are much more common than others and whether there are any values that appear to differ greatly from the others. These 'extreme' values are usually termed *outliers*.

To illustrate the histogram, let us consider data gathered in a radioactive decay experiment. In an experiment to study the emission of beta particles from a strontium 90 source, measurements were made of the number of particles emitted from the source over 100 consecutive 1 minute periods. The data gathered are shown in table 1.5. Inspection of the table indicates that all the values lie between about 1100 and 1400, but little else can be discerned. Do some values occur more often than others and if so which values? A good starting point for establishing the distribution of the data is to count the number of values (referred to as the *frequency*) which occur in predetermined intervals of equal width. The next step is to plot a graph consisting of frequency on the vertical

Table 1.5. *Counts from a radioactivity experiment.*

1265	1196	1277	1320	1248	1245	1271	1233	1231	1207
1240	1184	1247	1343	1311	1237	1255	1236	1197	1247
1301	1199	1244	1176	1223	1199	1211	1249	1257	1254
1264	1204	1199	1268	1290	1179	1168	1263	1270	1257
1265	1186	1326	1223	1231	1275	1265	1236	1241	1224
1255	1266	1223	1233	1265	1244	1237	1230	1258	1257
1252	1253	1246	1238	1207	1234	1261	1223	1234	1289
1216	1211	1362	1245	1265	1296	1260	1222	1199	1255
1227	1283	1258	1199	1296	1224	1243	1229	1187	1325
1235	1301	1272	1233	1327	1220	1255	1275	1289	1248

Table 1.6. *Grouped frequency distribution for data shown in table 1.5.*

Interval (counts)	Frequency
$1160 < x \leq 1180$	3
$1180 < x \leq 1200$	10
$1200 < x \leq 1220$	7
$1220 < x \leq 1240$	24
$1240 < x \leq 1260$	25
$1260 < x \leq 1280$	16
$1280 < x \leq 1300$	6
$1300 < x \leq 1320$	4
$1320 < x \leq 1340$	3
$1340 < x \leq 1360$	1
$1360 < x \leq 1380$	1

axis versus interval on the horizontal axis. In doing this we create a histogram.

Table 1.6, created using the data in table 1.5, shows the number of values which occur in consecutive intervals of 20 counts beginning with the interval 1160 to 1180 counts and extending to the interval 1360 to 1380 counts. This table is referred to as a *grouped frequency distribution*. The distribution of counts is shown in figure 1.1. We note that most values are clustered between 1220 and 1280 and that the distribution is almost symmetric, with the suggestion of a longer 'tail' at larger counts. Other methods by which univariate data can be displayed include stem and leaf plots and

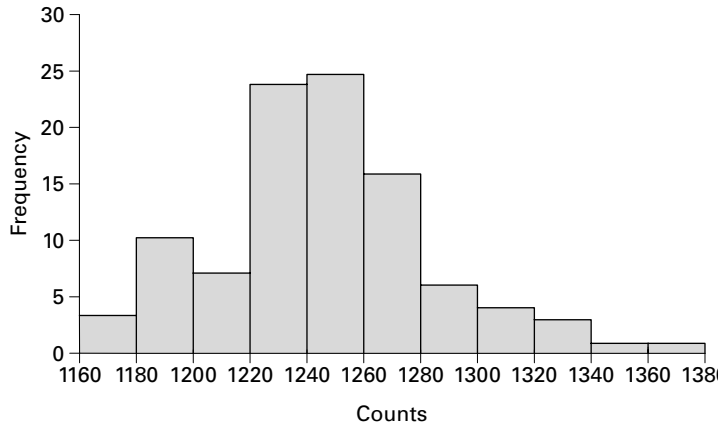

Figure 1.1. Histogram showing the frequency of counts in a radioactivity experiment.

pie charts,[8] though these tend to be used less often than the histogram in the physical sciences.

There are no 'hard and fast' rules about choosing the width of intervals for a histogram, but a good histogram:

- is easy to construct, so intervals are chosen to reduce the risk of mistakes when preparing a grouped frequency distribution. For example, an interval between 1160 and 1180 is preferable to one from (say) 1158 to 1178.
- reveals the distribution of the data clearly. If too many intervals are chosen then the number of values in each interval is small and the histogram appears 'flat' and featureless. At the other extreme, if the histogram consists of only two or three intervals, then all the values will lie in those intervals and the shape of the histogram reveals little.

In choosing the total number of intervals, N, a useful rule of thumb is to calculate N using

$$N = \sqrt{n} \tag{1.3}$$

where n is the number of values. Once N has been rounded to a whole number, the interval width, w, can be calculated using

$$w = \frac{\text{range}}{N} \tag{1.4}$$

[8] See Blaisdell (1998) for details of alternate methods of displaying univariate data.

where range is defined as

$$\text{range} = \text{maximum value} - \text{minimum value} \tag{1.5}$$

We should err on the side of selecting 'easy to work with' intervals, rather than holding rigidly to the value of w given by equation (1.4). If, for example, w were found using equation (1.4) to be 13.357, then a value of w of 10 or 15 should be considered, as this would make tallying up the number of values in each interval less prone to mistakes.

If there are many values then plotting a histogram 'by hand' becomes tedious. Happily, there are many computer based analysis packages, such as spreadsheets (discussed in chapter 2), which reduce the effort that would otherwise be required.

Exercise D

Table 1.7 shows the values of 52 'weights' of nominal mass 50 g used in an undergraduate laboratory. Using the values in table 1.7, construct

(i) a grouped frequency distribution;
(ii) a histogram.

Table 1.7. *Values of 52 weights.*

Mass (g)								
50.42	50.09	49.98	50.16	50.10	50.18	50.12	49.95	50.05
50.14	50.07	50.15	50.06	50.22	49.90	50.09	50.18	50.04
50.02	49.81	50.10	50.16	50.06	50.14	50.20	50.06	49.84
50.07	50.08	50.19	50.05	50.13	50.13	50.08	50.05	50.01
49.84	50.11	50.11	50.05	50.15	50.17	50.05	50.12	50.30
49.97	50.05	50.09	50.17	50.08	50.21	50.21		

1.4.2 Relationships and the *x*–*y* graph

A preoccupation of many scientists is to discover, and account for, the relationship between quantities. This fairly innocent statement conceals the fact that a complex and sometimes unpredictable interplay between experiment and theory is required before any relationship can be said to be accounted for in a quantitative as well as qualitative manner. Examples of relationships that may be studied through experiment are:

- the intensity of light emitted from a light emitting diode (LED) as the temperature of the LED is reduced;

- the power output of a solar cell as the angle of orientation of the cell with respect to the sun is altered;
- the change in electrical resistance of a humidity sensor as the humidity is varied;
- the variation of voltage across a conducting ceramic as the current through it changes;
- the decrease in the acceleration caused by gravity with depth below the earth's surface.

Let us consider the last example in a little more detail, in which the free-fall acceleration caused by gravity varies with depth below the earth's surface. Based upon considerations of the gravitational attraction between bodies, it is possible to predict a relationship between acceleration and depth when a body has uniform density. By gathering 'real data' this prediction can be examined. Conflict between theory and experiment might suggest modifications are required to the theory or perhaps indicate that some 'real' anomaly, such as the existence of large deposits of gold close to the site of the measurements, has influenced the values of acceleration.

As the acceleration in the example above depends on depth, we refer to the acceleration as the *dependent* variable, and the depth as the *independent* variable. (The independent and dependent variables are sometimes referred to as the predictor and response variables respectively.) A convenient way to record values of the dependent and independent variables is to construct a table. Though concise, a table of data is fairly dull and cannot assist efficiently with the identification of trends or patterns in data. A revealing and very popular way to display bivariate data is to plot an x–y graph (sometimes referred to as a scatter graph). The 'x' and the 'y' are the symbols used to identify the horizontal and vertical axes respectively of a Cartesian co-ordinate system.[9]

If properly prepared, a graph is a potent summary of many aspects of an experiment.[10] It can reveal:

- the quantities being investigated;
- the number and range of values obtained;
- gaps in the measurements;
- a trend between the x and y quantities;
- values that conflict with the trend shown by the majority of the data;
- the extent of uncertainty in the values (sometimes indicated by 'error bars').

[9] The horizontal and vertical axes are sometimes referred to as the abscissa and ordinate respectively.

[10] Cleveland (1994) discusses what makes 'good practice' in graph plotting.

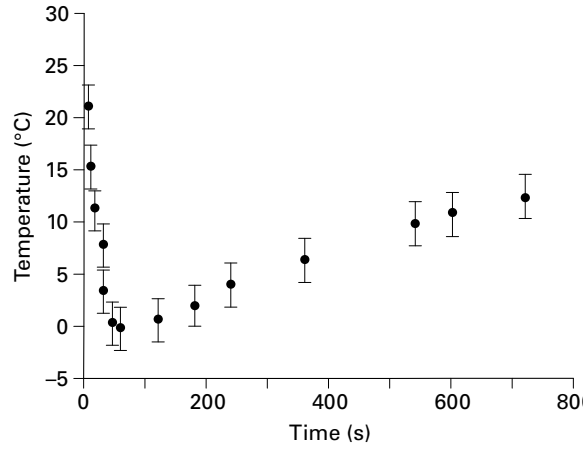

Figure 1.2. Temperature versus time for a thermoelectric cooler.

If a graph is well constructed, this qualitative information can be 'absorbed' in a few seconds. To construct a good graph we should ensure that:

- a caption describing the graph is included;
- axes are clearly labelled (and the label includes the unit of measurement);
- the scales for each axis are chosen so that plotting, if done by hand, is made easy so that values can be read easily from the graph;
- the graph is large enough to allow for the efficient extraction of information 'by eye';
- plotted values are clearly marked.

An *x*–*y* graph is shown in figure 1.2 constructed from data gathered in an experiment to establish the cooling capabilities of a thermoelectric cooler (TEC).[11] Attached to each point in figure 1.2 are lines which extend above and below the point. These lines are generally referred to as *error bars* and in this example are used to indicate the uncertainty in the values of temperature.[12] The '*y*' error bars attached to the points in figure 1.2 indicate that the uncertainty in the temperature values is about 2 °C. As '*x*' error bars are absent we infer that the uncertainty in values of time is too small to plot on this scale.

[11] A thermoelectric cooler is a device containing junctions of semiconductor material. When a current passes through the device, some of the junctions expel thermal energy (causing a temperature rise) while others absorb thermal energy (causing a temperature drop).

[12] Chapter 5 considers uncertainties in detail.

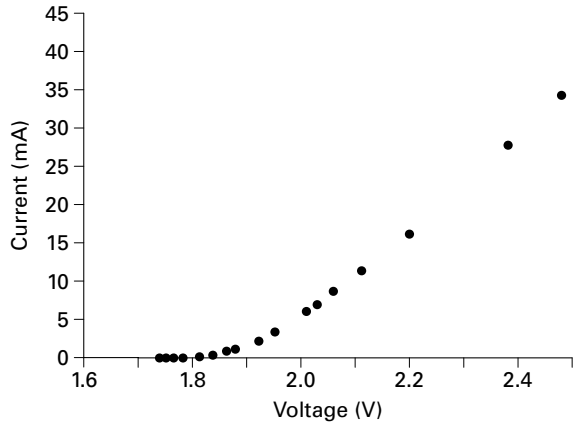

Figure 1.3. Variation of current with voltage for a LED.

If an x–y graph consists of many points, a convenient way to plot those points and attach error bars is to use a computer based spreadsheet (see section 2.7.1).

1.4.3 Logarithmic scales

The scales on the graph in figure 1.2 are linear. That is, each division on the x axis corresponds to a time interval of 200 s and each division on the y axis corresponds to a temperature interval of 5 °C. In some situations important information can be obscured if linear scales are employed. As an example, consider the current–voltage relationship for a LED as shown in figure 1.3. It is difficult to determine the relationship between current and voltage for the LED in figure 1.3 for values of voltage below about 2 V. As the current data span several orders of magnitude, the distribution of values can be more clearly discerned by replacing the linear y scale in figure 1.3 by a logarithmic scale. Though graph paper is available in which the scales are logarithmic, many computer based graph plotting routines, including those supplied with spreadsheet packages, allow easy conversion of the y or x or both axes from linear to logarithmic scales. Figure 1.4 shows the data from figure 1.3 replotted using a logarithmic y scale. As one of the axes remains linear, this type of graph is sometimes referred to as semi-logarithmic.

Exercise E

The variation of current through a Schottky diode is measured as the temperature of the diode increases. Table 1.8 shows the data gathered in the experiment. Choosing appropriate scales, plot a graph of current versus temperature for the Schottky diode.

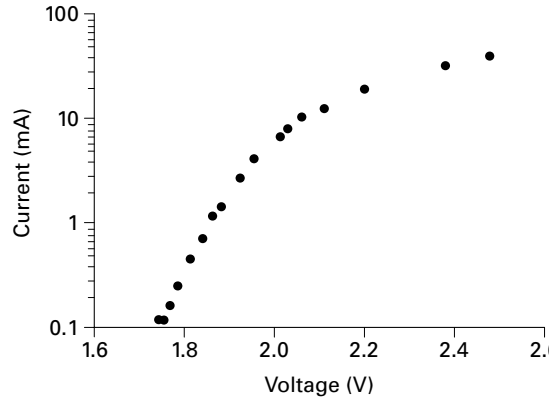

Figure 1.4. Current versus voltage using semi-logarithmic scales on the x and y axes.

Table 1.8. *Variation of current with temperature for a Schottky diode.*

Temperature (K)	Current (A)
297	2.86×10^{-9}
317	1.72×10^{-8}
336	6.55×10^{-8}
353	2.15×10^{-7}
377	1.19×10^{-6}
397	3.22×10^{-6}
422	1.29×10^{-5}
436	2.45×10^{-5}
467	9.97×10^{-5}
475	1.41×10^{-4}

1.5 Key numbers summarise experimental data

A significant challenge facing all experimenters is to find ways to express data in a concise fashion without obscuring important features. The histogram can give us the 'big picture' regarding the distribution of values and can alert us to important features such as lack of symmetry in the distribution, or the existence of outliers. This information, though very important, is essentially qualitative. What quantitative measures can we use to summarise all the data?